

Introduction to Bioinformatics

5. Multiple Sequence Alignment and Phylogenetic Trees

Benjamin F. Matthews

United States Department of Agriculture
Soybean Genomics and Improvement
Laboratory

Beltsville, MD 20708

matthewb@ba.ars.usda.gov

What we will cover today

- Multiple Sequence Alignment (MAS)
- Motifs
- Phylogenetic Trees

What is Multiple Sequence Alignment (MSA)?

- An extension of a pairwise alignment
- Can be local or global
- The “inputs” are the same
 - A set of amino acid or nucleic sequences
 - Substitution (scoring) matrices
 - Gap penalties
- The objectives are similar: find an alignment of more than two sequences
- Discussed in earlier lecture

Multiple Sequence Alignment

Wheat MSADKPSAYMLWLSNARESIKRENPDSGIL
Rice MKADKPSAYML - - - NARESI- - ENPDSGRL
Soy MPADKPSMFML - - - NPSESI - - NPDSARL

Why Do Multiple Sequence Alignment?

- Characterize protein families by identifying shared regions of homology
- Determine the consensus sequence of several aligned sequences
- Establish relationships and phylogenies
 - Clustering analysis
 - Structural modeling
 - Evolutionary analysis
- Use in a database search of protein families

Multiple Alignment programs

align several protein sequences

-
- ClustalW
 - <http://www.ch.embnet.org/software/ClustalW.html>
 - Multiple sequence alignment program
 - T-Coffee
 - http://igs-server.cnrs-mrs.fr/~cnotred/Projects_home_page/t_coffee_home_page.html
 - Alignment program that often gives better results, especially when dealing with divergent sequences and long insertions

Multiple sequence alignment

- ClustalW
 - Does alignment and phylogenetic tree
 - www.ebi.ac.uk/clustalw/index.html
- Dialign
 - Bibiserv.techfak.uni-bielefeld.de/dialign/
- Tcoffee
 - Igs-seerver.cnrs-mrs.fr/Tcoffee

MSA Algorithms

- As with the pairwise sequence comparisons, there are two types of multiple alignment algorithms
 - Optimal
 - Heuristic

Optimal MSA

- Extension of dynamic programming to multiple dimensions
- Exhaustive search
- Guaranteed to find an optimal score
- Need an n-dimensional matrix for scoring
- Computationally expensive
- Time complexity for pairwise comparisons is $O(m_1 * m_2)$; for multiple alignment should be $O(m^n)$
- Not feasible for $n > 10$ sequences of length $m > 200$ residues

Heuristic MSA

- Limit the exhaustive search
- Attempt to rapidly find a good, but not necessarily optimal alignment
- Most popular methods:
 - Tree alignments
 - Star alignments

Heuristic approaches to MSA

- **Progressive global alignment starting from the most similar sequences:**
CLUSTALW
 - Pairwise alignment: calculate distance matrix
 - Neighbor joining: draw guide tree
 - Progressive alignment: align following guide tree
- **Iterative methods:** make initial crude alignment, then revise it: **DIALIGN**
- **Alignment based on locally conserved patterns found in sequences in the same order:** **BLOCKS, eMOTIF, GIBBS, MEME**
- **Use statistical methods and probabilistic models of the sequences:**
HMMER, SAM

What is a motif?

- **A short conserved region in DNA, RNA or protein sequence**
- **Corresponds to a structural or functional feature in proteins**
- **Shared by several sequences and can be generated by MSA**
- **Can be represented using position-specific scoring matrices**

What is a profile?

- A position-specific scoring matrix, or matrix of scores representing a motif
- 22 columns, one for each of the 20 amino acids, and 2 for the penalties of opening and extending gaps
- The rows of the profile: aligned amino acid residues of a group of sequences
- Residues with the highest scores define a consensus

What is a protein family?

- A set of evolutionarily related proteins
- Members of a protein family may range from very similar to quite diverse
- Often share domains. Domain is a part of a protein (greater than a motif) that can fold and carry out a function independently.

Motif- and domain-oriented databases

- Secondary databases, small compared to GenBank
- Contain representations of conserved sequences shared by a sequence family
- Are primarily used for annotation of unknown sequences
- Examples: Pfam, Blocks, PRINTS, Prodom, PROSITE

Motifs and conserved domains

Pfam

- **Protein FAMily (Pfam) is a large collection of multiple sequence alignments of sequence motifs or domains**
- Made up of two parts: Pfam-A and Pfam-B
- Pfam-A: curated database of gapped profiles
- Pfam-B: generated automatically from sequences taken from the Prodom database that do not overlap with Pfam-A
- Use a Hidden Markov Model (HMM) to define domains or to align a set of sequences

Blocks

- Multiple sequence alignments without gaps that were used to construct the BLOSUM substitution matrices
- Generated automatically
- Correspond to the most conserved regions of a protein
- Better used to identify protein sequence domains or families rather than identify motifs

PROSITE

- A database of sequence patterns (~motifs) associated with protein family membership
- Developed by largely manual process of seeking the patterns that best fit particular protein families
- Patterns may be useful in assigning distant homologs to sequence families
- PROSITE patterns are very short => may result in false positive occurrences in unrelated sequences

PRINTS

- A compendium of protein fingerprints
- A fingerprint is a group of conserved motifs used to characterize a protein family
- The motifs do not overlap (separated along a sequence)

Prodom

- **An automatically generated collection of protein domains**
- **Better described as a software tool to visualizes a protein's sequence domain structure**

Profile searches

- Numerical representations of multiple sequence alignments
- Depends upon patterns or motifs containing conserved residues
- Represent the common characteristics of a protein family
- Can find similarities among sequences with little or no sequence identity
- Allows for the analysis of distantly-related proteins

ProfileScan

- Search sequence against a collection of profiles and patterns
- Databases available
 - PROSITE profiles
 - PROSITE patterns
 - PfamA
 - PfamB
- <http://hits.isb-sib.ch/cgi-bin/PFSCAN>

Profile Construction

```
ADPHIIIVAVPG  
GCHTWIAAEPG  
GWHICIAEPG  
GWHILIGETP  
RPHWIIIVAVPG  
KPHWIIIAEPG  
KVGQLEIAEPG  
RPDTWIAEAPG  
ADPHIIIVAVPG  
ADPHIIIVAVPG  
GCHWVIAEPG  
HQQTIVVAATPG
```

- Which residues are seen at each position?
- What is the frequency of observed residues?
- Which positions are conserved?
- Where can gaps be introduced?

Position-Specific Scoring Table

| Cons | A | B | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | U | V | W | X | Y | Z |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|-----|-----|----|-----|-----|------|-----|-----|---|---|
| G | 17 | 18 | 0 | 10 | 14 | -22 | 31 | 0 | -9 | 12 | -15 | -8 | 15 | 30 | 9 | 6 | 18 | 14 | 1 | -35 | -22 | 11 | | |
| P | -60 | 0 | -60 | 0 | 0 | -60 | 0 | 0 | 0 | 0 | 0 | 0 | 23 | 2 | -2 | 12 | 11 | 17 | -31 | -8 | 3 | | | |
| H | 5 | 24 | -12 | 29 | 25 | -20 | 8 | 32 | -9 | 9 | -30 | -6 | 22 | 7 | 30 | 10 | 0 | 4 | -6 | -26 | -7 | 27 | | |
| I | -1 | -12 | 6 | -13 | -13 | 35 | -12 | -13 | 63 | -11 | 40 | 29 | -25 | -9 | -14 | -15 | -6 | 7 | 26 | -17 | 8 | -11 | | |
| V | 3 | -13 | 1 | -31 | -8 | 22 | -3 | -11 | 66 | -9 | 37 | 30 | -13 | -3 | -9 | -13 | -6 | 6 | 50 | -19 | 2 | -8 | | |
| Y | 5 | -8 | 9 | -8 | 0 | 19 | -1 | -13 | 87 | -9 | 38 | 38 | -13 | -2 | -11 | -13 | -6 | 9 | 50 | -29 | 0 | -8 | | |
| S | 54 | 18 | 12 | 20 | 17 | -26 | 46 | -6 | -4 | -1 | -13 | -8 | 12 | 19 | 9 | -13 | 21 | 18 | 9 | -39 | -20 | 10 | | |
| W | 40 | 20 | 20 | 20 | 20 | -36 | 40 | -10 | 20 | 20 | -10 | 0 | 26 | 30 | -10 | -10 | 30 | 150 | 20 | -66 | -30 | 20 | | |
| R | 31 | 4 | 1 | 4 | 4 | 4 | 38 | 11 | 0 | 5 | 15 | 13 | 20 | 17 | 17 | 24 | 22 | 8 | -60 | -48 | 12 | | | |
| G | -60 | -22 | -22 | 70 | 50 | → | 150 | -30 | -30 | -10 | -80 | -30 | 40 | 30 | 20 | -30 | 60 | 40 | 20 | -100 | -70 | 30 | | |

Patterns

[FY] - x - C - x (2) - [VA] - x - H (3)

reads as:

| | |
|---------------------|-------------|
| Phe <i>or</i> Tyr | followed by |
| any amino acid | followed by |
| Cys | followed by |
| any two amino acids | followed by |
| Val <i>or</i> Ala | followed by |
| any amino acid | followed by |
| three His | |

ProfileScan

- <http://hits.isb-sib.ch/cgi-bin/PFSCAN>
- Find all known motifs in a sequence

Protein sequence

```
L A Q N P R S T L T P K A R G F S R L L  
Q I P E M A S V S A L A K Y K L V F L G  
D Q S V G K T S I I T R F M Y D K F D N  
T Y Q A T I G I D F L S K T M Y L E D R  
T V R L Q L W D T A G Q E R F R S L I P  
S Y I R D S S V A V I V Y D V A S R Q T  
F L N T A K W I E E V R T E R G S D V I  
I V L V G N K T D L V E K R Q V S I E E  
G E A K A R E L N V M F I E T S A K A G  
F N I K A L F R K I A A A L P G M E T L  
S S A K Q E D M V D V N L K S T N G S A  
Q S Q P Q S S G C A C *
```

Motif or conserved domain searches

- A pattern retained through evolution
 - not randomly changed by mutation
- Retained to help perform a specific function
- Domains can be found in databases
 - SMART
<http://smart.embl-heidelberg.de/>
 - Pfam
<http://www.sanger.ac.uk/Software/Pfam/>
 - COGs Clusters of Orthologous Groups
<http://www.ncbi.nlm.nih.gov/COG/>

Motif Scan in a Protein Sequence
(ProfileScan Server)

Motif scanning means finding all known motifs that occur in a sequence. This form lets you paste a protein sequence, select the collections of motifs to scan for, and launch the search. Some general documentation is available about the Prosite and Pfam collections of motifs. Another document deals with the interpretation of the match scores. You should consult the home pages of [Prosite](#) or ExPASy, [Pfam](#) and [InterPro](#) for additional information.

A pre-compiled list of matches is also available on our server ([Hits](#)). If your proteins of interest are already in the databases, the [Query by Protein](#) form is much faster, and the [Protein Hub](#) provides you a collection of tools that you might find useful.

| | | |
|--|--|--|
| Protein Sequence Input | <pre>L A Q N P R S T L T P K A R G F S R L L Q I P E M A S V S A L A K Y K L V F L G D Q S V G K T S I I T R F M Y D K F D N</pre> | <input type="button" value="Reset"/> <input type="button" value="Clear"/> |
| Motif Scan Parameters | | |
| <input checked="" type="checkbox"/> The Prosite profiles including the pre-released ones <input checked="" type="checkbox"/> The Prosite patterns <input type="checkbox"/> Database of motifs <input type="checkbox"/> Prosite patterns that match most frequently | | |

Motif Scan in a Protein Sequence - Microsoft Internet Explorer

Motif Scan in a Protein Sequence
result

Query

- Protein sequence:

```
>RAW_SEQUENCE
LAQNPRSTLTPKARGFSRLLQIPEMASVSALAKYKLVLFLGDQSVGKTSIITRFHYDKFDN
TYQATIGIDFLSKTMYLEDRTVRLQLWDTAQERFRSLIPSYIIRDSSVAVIVDVASROT
FLNTAKWIEVTERGSDVIIVLVGNKTDLVKEKRQVSIEGEAKARELNVMFIETSAKAG
FNIKALFRKIAAALPGMETLSSAKQEDMVDVNLLKSTNGSACSQPQSSGCACVVFIMCAPP
FCVIPFFVLCRFFSSTSLYEKMKEYISDRGRGYKLCLENFCC
```

- Databases: Prosite patterns (weekly-updated), Prosite profiles (weekly-update), Pfam collection of hidden Markov models (weekly-updated)

...scanning for weekly_pattern
...scanning for weekly_profile
...scanning for weekly_pfam
...confirming pfam

Result

- Summary:

Motif Scan in a Protein Sequence - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address: http://hits.isb-sib.ch/cgi-bin/PFSCAN_parser

Result

- Summary:

```
?pfam:ATP_BIND_1 pos. 38 - 196 E-value=0.73
?pfam:ARF pos. 21 - 194 E-value=3.3e-05
?pfam:GTP_EFTU pos. 36 - 204 E-value=4.1
! pfam:RAS pos. 35 - 196 E-value=3.3e-83
```

- Match Location:

```
query          LAQMPRSTLTPKARGTSRLLQIPPEMASVSALAKVYKLVLGDQSUGKTSIITRIMVVDKIDMTYQATIGIDFLSKTMVLEDR
pfam:ATP_BIND_1 <----->
pfam:ARF          <----->
pfam:GTP_EFTU      <----->
pfam:RAS          <----->

query          TURLOLMVTAGQERFRSLIPSVIRDSSVAVIVDVAASRGTFLNTRAKWIEEERTERGSDVU1VLUQHNTDLUKEPQVNSIEE
pfam:ATP_BIND_1 -----
pfam:ARF          -----
pfam:GTP_EFTU      -----
pfam:RAS          =====
```

```
query          GAKAKARELNUMTETSAKAGTWIKALPRKIAAHPGMETLSSAKQEDMUDQUNLKSTNG3AQ3QPQ35GCACUVTINMCAPP
pfam:ATP_BIND_1 <----->
pfam:ARF          <----->
pfam:GTP_EFTU      <----->
pfam:RAS          <----->
```

```
query          FCVIPVULCRPTTSSTSLYERKHEVYISDRGRGTYKLCLENFCC
```

- Prosite patterns (weekly-updated):

no match

- Prosite profiles (weekly-update):

Motif Scan in a Protein Sequence - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address: http://hits.isb-sib.ch/cgi-bin/PFSCAN_parser

Result

pos.: 36-204
raw-score = .90.8
N-score = 6.713
E-value = 4.1

Elongation_factor_Tu_GTP_binding_domain

Pfam-site
InterPro

165 SWAVIVYIVYASRGTFL...NTKAKTEEVTRGSDVIIILVLG

NtkDp

165 SWAVIVYIVYASRGTFL...NTKAKTEEVTRGSDVIIILVLG

208 NKTDLVEKROVS...IEEEENKAREL...

251 ...NMFIFLETSHGGGDNIMHLFRK...

294 ...IRHALPOMET-LSSHK+

35 ...KLVFLGQDQSVKTSITTRPMYOFDNYTQITGIDFLSKTMV...

78 EDRTVRLQDVTAGQERFRSLIPSYIROSSVIVYDWSRQT...

121 FLNTHKMEIEEVTRGSDVIIILVLGNTDL...

164 KROVSEEEGKARELN-WMFIETSKGGNIALFRKTAAL...

status: !
pos.: 35-196
raw-score = 289.9
N-score = 89.808
E-value = 3.3e-83

Ras family

Pfam-site
InterPro

Pfam: Search DNA vs. Pfam - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address: http://www.sanger.ac.uk/Software/Pfam/dnasearch.shtml

Pfam Protein families database of alignments and HMMs

Wellcome Trust Sanger Institute

Pfam: Search DNA vs. Pfam

Home Search by Browse by ftp iPFam Help

This form allows you to compare your DNA sequence against the whole of Pfam using the Wise2 software package

Cut and Paste your DNA sequence here, fasta format

```
ATATGCTTCGTTACATTACCCGTAAACCAAGTATTGTTCGT
TCCCTTTTGTGTTTACTCTGTGTGATATACTCTCAT
TATTCAAACTGTGAAGTGATATCATCATATTGCTCTAT
TTGGTGGTTAGGACCTGTTGTTGACCTCTGTCTAT
GTTTCAGAATTGTCATGTCAAAGTTGATTCTTATG
TTTTAAAG
```

* Searches now use the Sanger Blast queues. Presently the email option is not available *

It takes around 2 minutes for a 1,000 bp sequence, and around 2 hours for a 80kb sequence, depending on how many matches you get in them and the load on the sanger centre system

Or: Select the sequence file you wish to use

DNA sequence is Human Genomic DNA

Submit Query Reset Example

Output options for alignments

- Parameters
- Pretty alignment
- Predicted peptide
- Summary output
- GFF output
- Parseable alignment output

Output for gene predictions

- Gene structure as readable ASCII format
- EMBL feature table format
- EMBL feature table format suitable for Artemis
- AceDB subsequence object
- Translation as fasta format protein
- cDNA as fasta format DNA

If you think there is something wrong with this form or its results, please email Ewan Birney

Pfam: p450 - Netscape

http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00067

Pfam: p450

Protein families database of alignments and HMMs

Wellcome Trust Sanger Institute

Home Keyword Search Protein Search Browse Pfam DNA Search Taxonomy ftp Help p450 domain

Accession number: PF00067

Cytochrome P450

Add Annotation

Cytochrome P450s are involved in the oxidative degradation of various compounds. Particularly well known for their role in the degradation of environmental toxins and mutagens. Structure is mostly alpha, and binds a heme cofactor.

INTERPRO description (entry IPR001128)

The cytochrome P450 enzymes constitute a superfamily of haem-thiolate proteins. P450 enzymes usually act as terminal oxidases in multicomponent electron transfer chains, called P450-containing monooxygenase systems and are involved in metabolism of a plethora of both exogenous and endogenous compounds. P450-containing monooxygenase systems primarily fall into two major classes: bacterial/mitochondrial (type I), and microsomal (type II). All P450 enzymes can be categorised into two main groups, the so-called C- and E-classes. P450 proteins of prokaryotic 3-component systems and fungal P450nor (CYP55) belong to the C-class; all other known P450 proteins from distinct systems are of the E-class [MEDLINE:93135827].

QuickGO

PROCESS : Electron transport (GO:0006116)

Alignments
Domain Structure

The Swissprot/PDB mapping was provided by MSD

For additional annotation, see the PROSITE document 2DOC00081 [ExPasy] SRS-UK | SRS-USA

Constructing Phylogenetic Trees

Why use phylogenetics

- Determine closest relative of an organism
- Discover the function of a gene
 - Identify orthologous, well-characterized gene in another species
- Retrace the origin of a gene
 - Mutations, deletions, gene duplications, gain- or loss-of-function, inactivation

Important

- Data quality
 - Highly accurate multiple sequence alignment that contains properly chosen sequences

Types of genes

- Orthologs
 - Separated only by speciation
 - A common ancestor gives birth to two subgroups that slowly drift away to become distinct species
- Paralogs
 - A gene is duplicated. The resultant two genes slowly diverge in sequence
- Xenologs
 - Result from a lateral transfer of a gene from one organism into the genome of another organism

How do you know two genes
coming from two different
species are orthologs or paralogs?

No simple solution

Strategy

- 1) Choose a sequence from genome A
- 2) BLAST search sequence A against every sequence in complete genome of B
- 3) BLAST search returns sequence B as a top hit
- 4) BLAST search B against every sequence in genome A
- Is sequence B the top hit???
- This does not prove – but provides support for

Tips

- Avoid sequence fragments
- Avoid xenologs
- Avoid recombinant sequences (especially seen in viruses)
- Avoid very large complex families containing repeats
- Keep your data set small
- Add an out group to root your tree
 - ie. a sequence that you know is a member, but has diverged long ago from the rest of the set

How to improve your multiple sequence alignment for phylogenetics

- Remove gaps
 - Gaps cause problems

| | | | | | | |
|-------|-----------------|------------|------------|--------|----|-----------|
| Wheat | MSADKPSAYMLWLSN | [redacted] | [redacted] | ARES | I | KRENPDGIL |
| Rice | MKADKPSAYML | --- | N | ARESI | - | ENPDSGRL |
| Soy | MPADKPSMFML | --- | N | NPSESI | -- | NPDSARL |

How to improve your multiple sequence alignment for phylogenetics

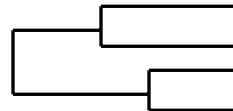
- Remove extremities of your multiple sequence alignment
- N-terminus and C-terminus tend to be poorly conserved
- Remove gap-rich regions
- Keep most informative blocks
 - Typically 20 to 30 amino acids long
 - Contains a few conserved positions

Tree software

- ClustalW
 - easy
- Phylogenetic
 - sophisticated

Example: tree alignment of four sequences

A _____
B _____
C _____
D _____



- Compare all six pairs of sequences
- Define and compute distances between the sequences
- Then use cluster analysis
- The number of pairs of N segments = $N(N-1)/2=4(3)/2=6$

Clustal W

- Format is important
- Can past or upload sequences
- Each sequence must have a unique name
- No empty lines
- No white spaces
- No control characters
- Limited to 500 sequences or 10MD, which ever is smaller)

```

>Arabidopsis
MVMAGASSLDEIRQAQRADGPAGILAIGTANPENHVLQAEYHDYYFRITNSEHKT
DLKEKFKRMCDKSTIRKRHMHLTEEFLKENPHMCAYMAPSLTRQDIVVVEPKL
GKEAAVKAIKEWGQPKSITHVVFCTTSVGDMPGADYQLTKLLGLRPSVKRLMM
YQQGCFAGGTVLRIAKDLAENNRRGARVLVVCSEITAVTFRGSPSDTHLDLSLVQAL
FSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTLPDSDGAIDGHLREVGLTFHLLKD
VPLGISKNIVKSLSDEAFKPLGISDWNSLFWIAHPGGPAILDQVEIKLGLKEEKMRAT
RHVLSEYGNMSSACVLFILDEMRRKSAKDGVATTGEGLEWGVSGFGPGLSvetv
VLHSVPL
>soyCH5
MVSVEEIRQAQRAGPATVMAIGTATPPNCVDQSTYPDYYFRITNSEHMTTELKEK
FKRMCDKSMIKKRYMLNEEILKENPSVCAYMAPSLDARQDMVVMEVPKLGKEA
ATKAIKEWGQPKSITHLIFCTTSVGDMPGADYQLTKLLGLRPSVKRYMMYQQGC
FAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPTDTHLDLSLVQALFGDGA
AAVIVGSDPLPVEKPLFQLVWTAQTLPDSEGAIDGHLREVGLTFHLLKDVGPLISK
NIEKALVEAFQPLGSDYNISFWIAHPGGPAILDQVEAKLGLKPEKMEATRHLSEY
GNMSSACVLFILDQMRKKSIENGLGTTGEGLDWGVLFVFGPGLTVETVVLRSVT
>SoyCh6
MVSVEEIRKAQRAGPATVMAIGTATPPNCVDQSTYPDYYFRITNSDHMNELKEKF
KRMCDSMIKKRYMLNEEILKENPSVCAYMAPSLDARQDMVVVEVPKLGKEAA
TKAIKEWGQPKSITHLIFCTTSVGDMPGADYQLTKLLGLRPSVKRYMMYQQGCF
AGGTVLRLAKDLAENNTGARVLVVCSEITAVTFRGSPSDTHLDLSLVQALFGDGAA
AVIVGSDPLPAEKPLFELVWTAQTLPDSEGAIDGHLREVGLTFHLLKDVGPLISKNI
QKALVEAFQPLGIDDYNSFWIAHPGGPAILDQVEAKLGLKPEKMEATRHLSEYG
NMSSACVLFILDQMRKKSIENGLGTTGEGLDWGVLFVFGPGLTVETVVLRSVT

```

ClustalW - MICROSOFT INTERNET EXPLORER

Help Index General Help Formats Gaps Matrix References ClustalW Help ClustalW FAQ Jalview Help Scores Table Alignment Guide Tree Colours

Address: http://www.ebi.ac.uk/clustalw/index.html#

SEQUENCE ANALYSIS

ClustalW Submission Form

Clustal W is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylogenograms. [New users, please read the FAQ.](#)

[Download Software](#)

| | | | | |
|--------------------------------------|--------------------------------------|--|---------------------------------------|---|
| YOUR EMAIL | ALIGNMENT TITLE | RESULTS | ALIGNMENT | CPU MODE |
| <input type="text"/> | Sequence | interactive <input type="button" value="▼"/> | full <input type="button" value="▼"/> | single <input type="button" value="▼"/> |
| KTUP (WORD SIZE) | WINDOW LENGTH | SCORE TYPE | TOPDIAG | PARGAP |
| def <input type="button" value="▼"/> | def <input type="button" value="▼"/> | percent <input type="button" value="▼"/> | def <input type="button" value="▼"/> | def <input type="button" value="▼"/> |
| MATRIX | GAP OPEN | END GAPS | GAP EXTENSION | GAP DISTANCES |
| def <input type="button" value="▼"/> | def <input type="button" value="▼"/> | def <input type="button" value="▼"/> | def <input type="button" value="▼"/> | def <input type="button" value="▼"/> |

| | | | | |
|--|--|---------------------------------------|--------------------------------------|--------------------------------------|
| OUTPUT | | PHYLOGENETIC TREE | | |
| OUTPUT FORMAT | OUTPUT ORDER | TREE TYPE | CORRECT DIST. | IGNORE GAPS |
| aln w/numbers <input type="button" value="▼"/> | aligned <input type="button" value="▼"/> | none <input type="button" value="▼"/> | off <input type="button" value="▼"/> | off <input type="button" value="▼"/> |

Enter or Paste a set of Sequences in any supported format:

Help

Internet

ClustalW - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address: http://www.ebi.ac.uk/clustalw/index.html#

| | | | | |
|-------------|----------|----------|---------------|---------------|
| (WORD SIZE) | LENGTH | | | |
| def | def | percent | def | def |
| MATRIX | GAP OPEN | END GAPS | GAP EXTENSION | GAP DISTANCES |
| def | def | def | def | def |

| OUTPUT | | PHYLOGENETIC TREE | | |
|---------------|--------------|-------------------|---------------|-------------|
| OUTPUT FORMAT | OUTPUT ORDER | TREE TYPE | CORRECT DIST. | IGNORE GAPS |
| aln w/numbers | aligned | none | off | off |

Enter or Paste a set of Sequences in any supported format: [Help](#)

```
>SoyCh6
MVSVEEIRKAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNS
DHNNELKEFKKMDCKSMIKKRVMYLNEEILKENPSCVAMEPSDL
ARQDNVVVEVPKLGKEAATKAIKEWQPKSKITHLIFCTTSGVDMR
GADYQLTKLLGLRPSVKRYMMYQQCCFAGGTVLRLAKDLAENNTGA
RVLVVCSEITAVTFRGPSDTLHDSLVGQALFGDGAAVIVGSDPLP
AEKPLFELVUTAQTLIPLDSEGAIIDGHLEVGFTFHLLKDVPGLISK
NIQKALVEAFQPLGIDDVNSIFIWAHPGGFAILDQVEAKLGLKPEK
MEATHVULSEYGNHSS&CVLFILDQMREKKSIENGLGTTGEGLENGV
LFGFGPGLTVETVVLSRVTV
```

Upload a file: [Browse...](#) [Run](#) [Reset](#)

If you plan to use these services during a course please contact us using the email below.
Please read the [FAQ](#) before seeking help from our support staff.

Internet

ClustalW - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address: http://www.ebi.ac.uk/cgi-bin/clustalw/result?tool=clustalw&jobid=clustalw-20050202-18582521&poll=yes

SEQUENCE ANALYSIS

Help General Help Formats Gaps Matrix References ClustalW Help ClustalW FAQ JalView Help Scores Table Alignment Guide Tree Colours

ClustalW Results

| Results of search | |
|---------------------|---|
| Number of sequences | 3 |
| Alignment score | 6306 |
| Sequence format | Pearson |
| Sequence type | aa |
| ClustalW version | 1.82 |
| JalView | JalView |
| Output file | clustalw-20050202-18582521.output |
| Alignment file | clustalw-20050202-18582521.aln |
| Guide tree file | clustalw-20050202-18582521.dnd |
| Your input file | clustalw-20050202-18582521.input |

[SUBMIT ANOTHER JOB](#)

To save a result file right-click the file link in the above table and choose "Save Target As".
If you cannot see the JalView button, reload the page and check your browser settings to enable Java Applets.

Applet jalview.ButtonAlignApplet started

Internet

http://www.ebi.ac.uk/cgi-bin/jobresults/clustalw/clustalw-20050202-18582521.aln - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites Home Print Copy Paste

Address http://www.ebi.ac.uk/cgi-bin/jobresults/clustalw/clustalw-20050202-18582521.aln Go Links

LUSTAL W (1.82) multiple sequence alignment

```

oyCH6      -----MVSVEIRKAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSDHNEELKEK 55
oyCH5      MVSVEIRQAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSEHTELKEK 55
rabidopsis MVMAGASSLDIEIRQAQRAQDGPAFILAIAGTANPENHVLQAEYHDYYFRITNSEHTELKEK 60
               *;*****:***:*** :;*****.* * * : * *****:***:***:****

oyCH6      FKRCMDKSMIKRKYMLNEEILKENPSVCAYMEPSLDAQDMVVVEVPKLGEKEAAKAIK 115
oyCH5      FKRCMDKSMIKRKYMLNEEILKENPSVCAYMAPSLDAQDMVVMEVPKLGEKEAAKAIK 115
rabidopsis FKRCMDKSTIRKRHMHLTEFLKENPHMCAYMAPSLDTRQDIVVVEVPKLGEAAVKAIAK 120
               *****:***:***:***:***:*****:*****:*****:*****:****

oyCH6      EWGQPKSKITHLIFCTTSGVMPGADYQLTKLLGLRPSVKRYMMYQQGCFAAGGTVLRLAK 175
oyCH5      EWGQPKSKITHLIFCTTSGVMPGADYQLTKLLGLRPSVKRYMMYQQGCFAAGGTVLRLAK 175
rabidopsis EWGQPKSKITHRVFCCTTSGVMPGADYQLTKLLGLRPSVKRLMMYQQGCFAAGGTVLRLAK 180
               *****:*****:*****:*****:*****:*****:*****:*****:*****:****

oyCH6      DLAEENNNTGARVLVVCSEITAVTFRGPDSDTHLDLSLVGQALFGDAAAIVGSDP--LPAEK 233
oyCH5      DLAEENNKGARVLVVCSEITAVTFRGPTDTHLDLSLVGQALFGDAAAIVGSDP--LPVEK 233
rabidopsis DLAEENNNGARVLVVCSEITAVTFRGPDSDTHLDLSLVGQALFGDAAAIVGSDPDTSVGEK 240
               *****:*****:*****:*****:*****:*****:*****:*****:*****:****

oyCH6      PLFELVWTAQTILPDSSEGAIIDGHLEREVGLTFHLLKDVPGLISKNIQRALVEAFQPLGIDD 293
oyCH5      PLFQLVWTAQTILPDSSEGAIIDGHLEREVGLTFHLLKDVPGLISKNIQRALVEAFQPLGIDS 293
rabidopsis PIFEMVSAAQTILPDSDGAIDGHLEREVGLTFHLLKDVPGLISRNIVRSLEAFKPLGIDS 300
               *;*:***:*****:*****:*****:*****:*****:*****:*****:****

oyCH6      YNSIFWIAHPGGPAILDQVEAKLGLKPEKMEATHRVLSEYGNMSSACVLFILDQMRKKSI 353
oyCH5      YNSIFWIAHPGGPAILDQVEAKLGLKPEKMEATHRVLSEYGNMSSACVLFILDQMRKKSI 353
rabidopsis WNSJLPWIAHPGGPAILDQVEIKLGKREEKHRATRVLSEYGNMSSACVLFILDEMRRKSA 360
               :*****:*****:*****:*****:*****:*****:*****:*****:*****:****

oyCH6      ENGLGTTGEGLEWGVLFGFPGPLTETVVLRSVT 388
oyCH5      ENGLGTTGEGLDWGVLFGFPGPLTETVVLRSVT 388
rabidopsis ENGLGTTGEGLEWGVLFGFPGPLTETVVLRSVT 388
               :*****:*****:*****:*****:*****:*****:*****:*****:****

Done
```

Internet

stalW - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites Home Print Copy Paste

Address http://www.ebi.ac.uk/cgi-bin/clustalw/result?tool=clustalw&jobid=clustalw-20050202-19573904&poll=yes Go Links

Get Nucleotide sequences for Go Site search

EMBL-EBI
European Bioinformatics Institute

Home About EBI Research Services Toolbox Databases Downloads Submissions

SEQUENCE ANALYSIS

ClustalW Results

| Results of search | |
|---|-----------------------------------|
| Number of sequences | 3 |
| Sequence format | Clustal |
| Sequence type | aa |
| ClustalW version | 1.82 |
| Output file | clustalw-20050202-19573904.output |
| Phylo tree file | clustalw-20050202-19573904.ph |
| Your input file | clustalw-20050202-19573904.input |
| <input type="button" value="SUBMIT ANOTHER JOB"/> | |

ClustalW Output

Net ClustalTree started

Internet

- Phylogram
 - Branch length represent real distances
- Cladogram
 - Branches indicate only branching order

ClustalW - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Mail Print Favorites Links

Address http://www.ebi.ac.uk/cgi-bin/clustalw/result?tool=clustalw&jobid=clustalw-20050202-20225158&pol=yes Go Links

Phylip Tree

```
{
SoyCH6:0.01750,
soyCH5:0.01412,
Arabidopsis:0.18604);
```

Phylogram

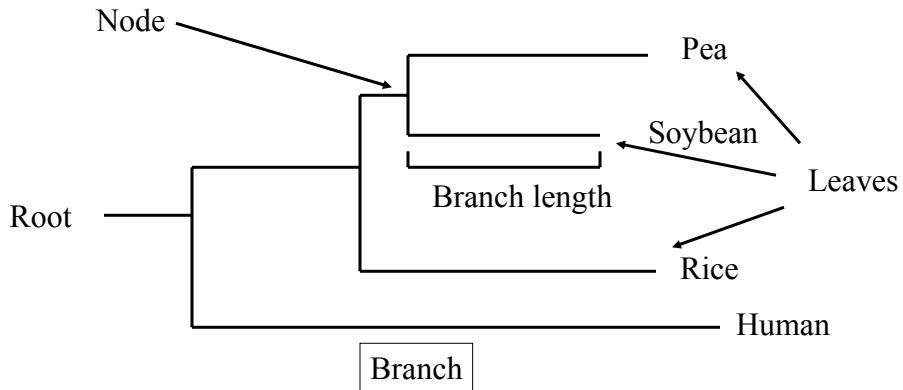
Right-click on the above tree to see display options.

Problems printing? Read [how to print a Phylogram or Cladogram](#).

Please contact [EBI Support](#) with any problems or suggestions regarding this site.
 [View Printer-friendly version of this page](#) | [Terms of Use](#)

Applet ClustalTree started

Parts of a Phylogenetic Tree



Tree building

- Trees are also called dendograms
- Nodes represent different organisms and links are used to show lines of descent
- Two basic types of questions about a tree:
 - a) its topology: how its interior nodes connect to one another and to the leaves
 - b) distance between pairs of nodes, which is an estimate of an evolutionary distance
- Tree may or may not have a root. A tree with root implies ancestral relationship between interior nodes

Phylogenetic tree evaluation

- How reliable phylogenetic three is?
- One criterion: if different methods of tree construction give the same result, this is good evidence that the tree is reliable
- Another criterion (bootstrapping): data are randomly sampled from any position within MSA, and are built into new artificial alignments, which are then tested by tree building
- Third criterion (jackknife): similar to bootstrapping, but instead of generating new datasets with replacement, it re-samples the original data set by dropping one or more alignment positions in each replicate

What we learned

- Multiple Sequence Alignment (MAS)
- Motifs
- Phylogenetic Trees